

Hierarchical Heavy Hitters with the Space Saving Algorithm

Michael Mitzenmacher Thomas Steinke
Justin Thaler

School of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138

Email: {jthaler, tsteinke}@seas.harvard.edu michaelm@eecs.harvard.edu

Abstract

The Hierarchical Heavy Hitters problem extends the notion of frequent items to data arranged in a hierarchy. This problem has applications to network traffic monitoring, anomaly detection, and DDoS detection. We present a new streaming approximation algorithm for computing Hierarchical Heavy Hitters that has several advantages over previous algorithms. It improves on the worst-case time and space bounds of earlier algorithms, is conceptually simple and substantially easier to implement, offers improved accuracy guarantees, is easily adopted to a distributed or parallel setting, and can be efficiently implemented in commodity hardware such as ternary content addressable memory (TCAMs). We present experimental results showing that for parameters of primary practical interest, our two-dimensional algorithm is superior to existing algorithms in terms of speed and accuracy, and competitive in terms of space, while our one-dimensional algorithm is also superior in terms of speed and accuracy for a more limited range of parameters.

1 Introduction

Finding *heavy hitters*, or frequent items, is a fundamental problem in the data streaming paradigm. As a practical motivation, network managers often wish to determine which IP addresses are sending or receiving the most traffic, in order to detect anomalous activity or optimize performance. Often, the large volume of network traffic makes it infeasible to store the relevant data in memory. Instead, we can use a *streaming* algorithm to compute (approximate) statistics in real time given sequential access to the data and using space sublinear in both the universe size and stream length.

We present and analyze a streaming approximation algorithm for a generalization of the Heavy Hitters problem, known as Hierarchical Heavy Hitters (HHHs). The definition of HHHs is motivated by the observation that some data are naturally hierarchical, and ignoring this when tracking frequent items may mean the loss of useful information. Returning to our example of IP addresses, suppose that a single entity controls all IP addresses of the subnet 021.132.145.*, where * is a wildcard byte. It is possible for the controlling entity to spread out traffic uniformly among this set of IP addresses, so that no single IP address within the set of addresses 021.132.145.* is a heavy hitter. Nonetheless, a network manager may want to know if the sum of the traffic of all IP addresses in the subnet exceeds a specified threshold.

One can expand the concept further to consider multidimensional hierarchical data. For example, one might track traffic between source-destination *pairs* of IP addresses at the router level. In that case, the network manager may want to know if there is a Heavy Hitter for network traffic at the level of two IP addresses, between a source IP address and a destination subnet, between a

Time Step	Update	Counter 1	Counter 2	Counter 3
0		<i>unused</i>	unused	unused
1	(a,+2)	(a,2)	<i>unused</i>	unused
2	(b,+6)	(a,2)	(b,6)	<i>unused</i>
3	(c,+4)	<i>(a,2)</i>	(b,6)	(c,4)
4	(a,+3)	(a,5)	(b,6)	<i>(c,4)</i>
5	(d,+4)	<i>(a,5)</i>	(b,6)	(d,8)
6	(e,+4)	(e,9)	<i>(b,6)</i>	(d,8)

Figure 1: Sample execution of Space Saving with 3 counters. Each counter tracks an item (denoted by a letter), and the estimated frequency of that item. The smallest counter is boldfaced and italicized.

source subnet and a destination IP address, or between two subnets. This motivates the study of the *two-dimensional* HHH problem.

There is some subtlety in the appropriate definitions, as it makes sense to require that an element is not marked as an HHH simply because it has a significant descendant, but because the *aggregation* of its children makes it significant; otherwise, the algorithm returns redundant, less helpful information. We present the definitions shortly, following previous work that has explored HHHs for both one-dimensional and multi-dimensional hierarchies [6, 12, 7, 8, 11, 15, 23].

HHHs have many applications, and have been central to proposals for real-time anomaly detection [25] and DDos detection [22]. While IP addresses serve as our motivating example throughout the paper, our algorithm applies to arbitrary hierarchical data such as geographic or temporal data. We demonstrate that our algorithm has several advantages, combining improved worst-case time and space bounds with more practical advantages such as simplicity, parallelizability, and superior performance on real-world data.

Our algorithm utilizes the Space Saving algorithm, proposed by Metwally et al. [18], as a subroutine. Space Saving is a *counter-based* algorithm for estimating item frequencies, meaning the algorithm tracks a subset of items from the universe, maintaining an approximate count for each item in the subset. Specifically, the algorithm input is a stream of pairs (i, c) where i is an item and $c > 0$ is a frequency increment for that item. At each time step the algorithm tracks a set T of items, each with a counter. If the next item i in the stream is in T , its counter is updated appropriately. Otherwise, the item with the smallest counter in T is removed and replaced by i , and the counter for i is set to the counter value of the item replaced, plus c . This approach for replacing items in the set may seem counterintuitive, as the item i may have an exaggerated count after placement, but the result is that if T is large enough, all Heavy Hitters will appear in the final set. Indeed, Space Saving has recently been identified as the most accurate and efficient algorithm in practice for computing Heavy Hitters [5], and, as we later discuss, it also possesses strong theoretical error guarantees [2].

1.1 Related Work

We require some notation to introduce prior related work and our contributions; this notation is more formally defined in Section 2. In what follows, N is the sum of all frequencies of items in the stream, ϵ is an accuracy parameter so that all outputs are within ϵN of their true count, and H represents the size of the hierarchy (specifically, the size of the underlying lattice) the data belongs to. Unitary updates refer to systems where the count for an item increases by only 1 on each step,

or equivalently, where we just count item appearances.

The one-dimensional HHH problem was first defined in [6], which also gave the first streaming algorithms for it. Several possible definitions and corresponding algorithms for the multi-dimensional problem were introduced in [7, 8]. The definition we use here is the most natural, and was considered in several subsequent works [12, 23]. In terms of practical applications, multi-dimensional HHHs were used in [10, 11] to find patterns of traffic termed “compressed traffic clusters”, in [25] for real-time anomaly detection, and in [22] for DDoS detection.

The Space Saving algorithm was used in [15] in algorithms for the one-dimensional HHH problem. Their algorithms require $O(H^2/\epsilon)$ space, while our algorithm requires $O(H/\epsilon)$ space. Very recently, [23] presented an algorithm for the two-dimensional HHH problem, requiring $O(H^{3/2}/\epsilon)$ space.

Other recent work studies the HHH problem with a focus on developing algorithms well-suited to commodity hardware such as ternary content-addressable memories (TCAMs) [13]. Our algorithms are also well-suited to commodity hardware, as we describe in Section 5. The primary difference between the present work and [13] is that the algorithms of [13] reduce overhead by only updating rules periodically, rather than on a per-packet basis. This leads to lightweight algorithms with *no* provable accuracy guarantees. However, simulation results in [13] suggest these algorithms perform well in practice. In contrast, our algorithms possess very strong accuracy guarantees, but likely result in more overhead than the approach of [13]. Which approach is preferable may depend on the setting and on the constraints of the data owner.

1.2 Our Contributions

In solving the Approximate HHH problem, there are three metrics that we seek to optimize: the time and space required to process each update and to output the list of approximate HHHs and their estimated frequencies and the quality of the output, in terms of the number of prefixes in the final output and the accuracy of the estimates. Our approach has several advantages over previous work.

1. The worst-case space bound of our algorithm is $O(H/\epsilon)$. Notice this does not depend on the sum of the item frequencies, N , as H depends only on the size of the underlying hierarchy and is independent of N . This beats the worst-case space bound of $O(\frac{H}{\epsilon} \log \epsilon N)$ from [7] and [8], the $O(H^2/\epsilon)$ bound for the one-dimensional algorithm of [15], and the $O(H^{3/2}/\epsilon)$ bound for the two-dimensional algorithm of [23]. Additionally our algorithm provably requires $o(H/\epsilon)$ space under realistic assumptions on the frequency distribution of the stream.
2. The worst-case time bound for our algorithm *per* insertion operation is $O(H \log \frac{1}{\epsilon})$ in the case of arbitrary updates and $O(H)$ in the case of unitary updates. Again this does not depend on N . Previous time bounds per insert were $O(H \log \epsilon N)$ in [6, 7, 8].
3. We obtain a refined analysis of error propagation to achieve better accuracy guarantees and provide non-trivial bounds on the number of HHHs output by our algorithm in one and two dimensions. These bounds were not provided for the algorithms in [6, 7, 8].
4. The space usage of our algorithm can be fixed *a priori*, independent of the sum of frequencies N , as it only depends on the number of counters maintained by each instance of Space Saving, which we set at $\frac{1}{\epsilon}$ in the absence of assumptions about the data distribution. In contrast, the space usage of the algorithms of [7] and [8] depends on the input stream, and these algorithms dynamically add and prune counters over the course of execution, which can be infeasible in realistic settings.

5. Our algorithm is conceptually simpler and substantially easier to implement than previous algorithms. We firmly believe *programmer time* should be viewed as a resource similar to running time and space. We were able to use an off-the-shelf implementation of Space Saving, but this fact notwithstanding, we still spent roughly an order of magnitude less time implementing our algorithms, compared to those from [7, 8].
6. Our algorithms extend easily to more restricted settings. For example, we describe in Section 5 how to efficiently implement our algorithms using TCAMs, how to parallelize them, how to apply them to distributed data streams, and how to handle sliding windows or streams with deletions.

We present experimental results showing that for parameters of primary practical interest, our two-dimensional algorithm is superior to existing algorithms in terms of speed and accuracy, and competitive in terms of space, while our one-dimensional algorithm is also superior in terms of speed and accuracy for a more limited range of parameters. In short, we believe our algorithm offers a significantly better combination of simplicity and efficiency than any existing algorithm.

2 Notation, Definitions, and Setup

2.1 Notation and Definitions

As mentioned above, the theoretical framework developed in this section was described in [8], and considered in several subsequent works [12, 23].

In examples throughout this paper, we consider the IP address hierarchy at bitwise granularity: for example, the *generalization* of 021.132.145.146 by one byte is 021.132.145.*, by two bytes is 021.132.*, by three bytes is 021.*, and by four bytes is *. In two dimensions, we consider pairs of IP addresses, corresponding to source and destination IPs. Each IP prefix that is not fully general in either dimensions has two parents. For example, the two parents of the IP pair (021.132.145.146, 123.122.121.120) are (021.132.145.*, 123.122.121.120) and (021.132.145.146, 123.122.121.*).

In general, let the dimension of our data be d , and the height of the hierarchy in the i 'th dimension be h_i . In the case of pairs of IP addresses, $d = 2$ and $h_1 = h_2 = 4$. Denote by $\text{par}(e, i)$ the generalization of element e on dimension i ; for example, if

$$e = (021.132.145.*, 123.122.121.120)$$

then $\text{par}(e, 1) = (021.132. *, 123.122.121.120)$ and $\text{par}(e, 2) = (021.132.145.*, 123.122.121.*)$. Denote the generalization relation by \prec ; for example,

$$(021.132.145.*, 123.122.121.120) \prec (021.132. *, 123.122. *.*)$$

Define $p \preceq q$ by $(p \prec q) \vee (p = q)$. The generalization relation defines a lattice structure in the obvious manner. We overload our notation to define the sublattice of a *set* of elements P as $(e \preceq P) \iff \exists p \in P$ such that $e \preceq p$. Let H denote the total number of nodes in the lattice: $H = \prod_{i=1}^d (h_i + 1)$.

We call an element *fully specified*, if it is not the generalization of any other element, e.g. 021.132.145.163 is fully specified. We call an element fully general in dimension i if $\text{par}(e, i)$ does not exist. We refer to the unique element that is fully general in all dimensions as the *root*. For ease of reference, we label each element in the lattice with a vector of length d , whose i 'th entry

is at most h_i , to indicate which lattice node the element belongs to, with the vector corresponding to each fully specified element having i 'th entry equal to h_i , and the vector corresponding to the root having all entries equal to 0. For example, the element (021.132.145.*, 123.122.121.120) is assigned vector (3, 4), and (021.*.*.*, 123.122.121.*) is assigned vector (1, 3). We define $\text{Level}(i)$ of the lattice to be the set of labels for which the sum of the entries in the label equals i . We overload terminology and refer to an element p as a member of $\text{Level}(i)$ if the label assigned to p is in $\text{Level}(i)$. Let $L = \sum_{i=1}^d h_i$ denote the deepest level in the hierarchy, that of the fully specified elements.

Definition 2.1. (*Heavy Hitters*) Given a multiset S of size N and a threshold ϕ , a Heavy Hitter (HH) is an element whose frequency in S is no smaller than ϕN . Let $f(e)$ denote the frequency of each element e in S . The set of heavy hitters is $HH = \{e : f(e) \geq \phi N\}$.

From here on, we assume we are given a multiset S of (fully-specified) elements from a (possibly multidimensional) hierarchical domain of depth L , and a threshold ϕ .

Definition 2.2. (*Unconditioned count*) Given a prefix p , define the unconditioned count of p as $f(p) = \sum_{e \in S \wedge e \preceq p} f(e)$.

The exact HHHs are defined inductively as the set of prefixes whose *conditioned count* exceeds ϕN , where the conditioned count is the sum of all descendant nodes that are neither HHHs themselves nor the descendant of an HHH. Formally:

Definition 2.3. (*Exact HHHs*) The set of exact Hierarchical Heavy Hitters are defined inductively.

1. \mathcal{HHH}_L , the hierarchical heavy hitters at level L , are the heavy hitters of S , that is the fully specified elements whose frequencies exceed ϕN .
2. Given a prefix p from $\text{Level}(l)$, $0 \leq l < L$, define \mathcal{HHH}_{l+1}^p to be the set $\{h \in \mathcal{HHH}_{l+1} \wedge h \prec p\}$ i.e. \mathcal{HHH}_{l+1}^p is the set of descendants of p that have been identified as HHHs. Define the conditioned count of p to be $F_p = \sum_{(e \in S) \wedge (e \preceq p) \wedge (e \not\preceq \mathcal{HHH}_{l+1}^p)} f(e)$. The set \mathcal{HHH}_l is defined as

$$\mathcal{HHH}_l = \mathcal{HHH}_{l+1} \cup \{p : (p \in \text{Level}(l) \wedge (F_p \geq \phi N))\}.$$

3. The set of exact Hierarchical Heavy Hitters \mathcal{HHH} is defined as the set \mathcal{HHH}_0 .

Figure 2 displays the exact HHHs for a two-dimensional hierarchy defined over an example stream.

Finding the set of hierarchical heavy hitters and estimating their frequencies requires linear space to solve exactly, which is prohibitive. Indeed, even finding the set of heavy hitters requires linear space [20], and the hierarchical problem is even more general. For this reason, we study the *approximate* HHH problem.

Definition 2.4. (*Approximate HHHs*) Given parameter ϵ , the Approximate Hierarchical Heavy Hitters problem with threshold ϕ is to output a set of items P from the lattice, and lower and upper bounds $f_{\min}(p)$ and $f_{\max}(p)$, such that they satisfy two properties, as follows.

1. *Accuracy.* $f_{\min}(p) \leq f(p) \leq f_{\max}(p)$, and $f_{\max}(p) - f_{\min}(p) \leq \epsilon N$ for all $p \in P$.
2. *Coverage.* For all prefixes p , define P_p to be the set $\{q \in P : q \prec p\}$. Define the conditioned count of p with respect to P to be $F_p = \sum_{(e \in S) \wedge (e \preceq p) \wedge (e \not\preceq P_p)} f(e)$. We require for all prefixes $p \notin P$, $F_p < \phi N$.

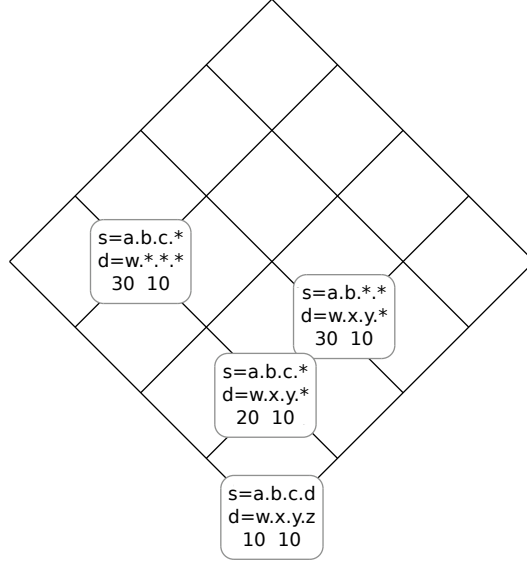


Figure 2: Example depicting exact HHHs for a two-dimensional stream of IP addresses at byte-wise granularity, using the threshold $\phi N = 10$. The exact HHHs consist of *ordered pairs* of source-destination IP-address prefixes (s denotes source and d denotes destination). Unconditioned counts of each HHH are on the left, and conditioned counts for each HHH are on the right. The stream consists of ten repetitions of the item $(a.b.c.d, w.x.y.z)$, followed by one instance each of items $(a.b.c.i, w.x.y.i)$, $(a.b.i.d, w.x.y.i)$, and $(a.b.c.i, w.i.y.z)$ for all i in the range 0 to 9. Here a, b, c, d, w, x, y , and z represent some distinct integers between 10 and 255.

Intuitively, the Approximate HHH problem requires outputting a set P such that no prefix with large conditioned count (with respect to P) is omitted, along with accurate estimates for the *unconditioned* counts of prefixes in P . One might consider it natural to require accurate estimates of the *conditioned* counts of each $p \in P$ as well, but as shown in [12], $\Omega(1/\phi^{d+1})$ space would be necessary if we required equally accurate estimates for the conditioned counts, and this can be excessively large in practice.

2.2 Our Algorithm, Sketched

Our algorithm utilizes the Space Saving algorithm, proposed by Metwally et al. [18] as a subroutine, so we briefly describe it and some of its relevant properties. As mentioned, Space Saving takes as input a stream of pairs (i, c) , where i is an item and $c > 0$ is a frequency increment for that item. It tracks a small subset T of items from the stream with a counter for each $i \in T$. If the next item i in the stream is in T , its counter is updated appropriately. Otherwise, the item with the smallest counter in T is removed and replaced by i , and the counter for i is set to the counter value of the item replaced, plus c . We now describe guarantees of Space Saving from [2].

Let N be the sum of all frequencies of items in the stream, let m be the number of counters maintained by Space Saving, and for any $j < m$, let $N^{\text{res}(j)}$ denote the sum of all but the top j frequencies. Berinde et al. [2] showed that for any $j < m$,

$$\forall i \left| f(i) - \hat{f}(i) \right| \leq \frac{N^{\text{res}(j)}}{m - j}, \quad (1)$$

where $\hat{f}(i)$ and $f(i)$ are the estimated and true frequencies of item i , respectively. By setting $j = 0$, this implies that $|f_i - \hat{f}_i| \leq \frac{N}{m}$, so only $\frac{1}{\epsilon}$ counters are needed to ensure error at most ϵN in any estimated frequency. For frequency distributions whose “tails” fall off sufficiently quickly, Space Saving provably requires $o(\frac{1}{\epsilon})$ space to ensure error at most ϵN (see [2] for more details).

Using a suitable min-heap based implementation of Space Saving, insertions take $O(\log m)$ time, and lookups require $O(1)$ time under arbitrary positive counter updates. When all updates are unitary (of the form $c = 1$), both insertions and lookups can be processed in $O(1)$ time using the *Stream Summary* data structure [18].

Our algorithm for HHH problems is conceptually simple: it keeps one instance of a Heavy Hitter algorithm at each node in the lattice, and for every update e we compute all generalizations of e and insert each one separately into a different Heavy Hitter data structure. When determining which prefixes to output as approximate HHHs, we start at the bottom level of the lattice and work towards the top, using the inclusion-exclusion principle to obtain estimates for the *conditioned* counts of each prefix. We output any prefix whose estimated conditioned count exceeds the threshold ϕN .

We mention that the ideas underlying our algorithm have been implicit in earlier work on HHHs, but have apparently been considered impractical or otherwise inferior to more complicated approaches. Notably, [8] briefly proposes an algorithm similar to ours based on *sketches*. Their algorithm can handle deletions as well as insertions, but it requires more space and has significantly less efficient output and insertion procedures. Significantly, this algorithm is only mentioned in [8] as an extension, and is not studied experimentally. An algorithm similar to ours is also briefly described in [12] to show the asymptotic tightness of a lower bound argument. Interestingly, they clearly state their algorithm is not meant to be practical. Finally, [6] describes a procedure similar to our one-dimensional algorithm, but concludes that it is both slower and less space efficient than other algorithms. We therefore consider one of our primary contributions to be the identification of our approach as not only practical, but in fact superior in many respects to previous more complicated approaches.

We chose the Space Saving algorithm [18] as our Heavy Hitter algorithm. In contrast, the algorithms of [7, 8] are conceptually based on the Lossy Counting Heavy Hitter algorithm [17]. A number of the advantages enjoyed by our algorithm can be traced directly to our choice of Space Saving over Lossy Counting, but not all. For example, the one-dimensional HHH algorithm of [15] is also based on Space Saving, yet our algorithm has better space guarantees.

3 One-Dimensional Hierarchies

We now provide pseudocode for our algorithm in the one-dimensional case, which is much simpler than the case of arbitrary dimension. As discussed, we use the Space Saving algorithm at each node of the hierarchy, updating all appropriate nodes for each stream element, and then conservatively estimate conditioned counts to determine an appropriate output set.

INITIALIZEHHH()

- 1 Initialize an instance $SS(n)$ of Space Saving with ϵ^{-1} counters at each node n of the hierarchy.

```

INSERTHHH(element  $e$ , count  $c$ )
1 /*Line 4 tells the  $n$ 'th instance of Space Saving
   to process  $c$  insertions of prefix  $p^*$ /
2 for all  $p$  such that  $e \preceq p$ 
3     Let  $n$  be the lattice node that  $p$  belongs to
4     UpdateSS( $SS(n)$ ,  $p$ ,  $c$ )

OUTPUTHHH1D(threshold  $\phi$ )
1 /*  $\text{par}(e)$  is parent of  $e^*$ /
2 Let  $s_e = 0$  for all  $e$ 
3 /* $s_e$  conservatively estimates the difference
   between unconditioned and conditioned counts of  $e^*$ /
4 for each  $e$  in postorder
5     ( $f_{\min}(e)$ ,  $f_{\max}(e)$ ) = GetEstimateSS( $SS(n)$ ,  $e$ )
6     if  $f_{\max}(e) - s_e \geq \phi N$ 
7         print( $e$ ,  $f_{\min}(e)$ ,  $f_{\max}(e)$ )
8          $s_{\text{par}(e)} += f_{\min}(e)$ 
9     else  $s_{\text{par}(e)} += s_e$ 

```

Figure 3 illustrates an execution of our one-dimensional algorithm on a stream of IP addresses at byte-wise granularity.

Time T	Counter 1	Counter 2	Counter 3
Level 1	(a.*.*., 20)	(<i>h.*.*., 12</i>)	(q.*.*., 16)
Level 2	(a.b.*.*., 18)	(<i>h.i.*.*., 12</i>)	(q.r.*.*., 16)
Level 3	(a.b.c.*., 18)	(<i>q.r.s.*., 9</i>)	(h.i.j.*., 14)
Level 4	(a.b.c.d., 10)	(h.i.m.n., 15)	(<i>a.b.c.e., 8</i>)
(w.x.y.z., +3)			
Time T+1	Counter 1	Counter 2	Counter 3
Level 1	(a.*.*.*., 20)	(<i>w.*.*.*., 15</i>)	(q.*.*.*., 16)
Level 2	(a.b.*.*.*., 18)	(<i>w.x.*.*.*., 15</i>)	(q.r.*.*.*., 16)
Level 3	(a.b.c.*.*., 18)	(<i>w.x.y.*., 12</i>)	(h.i.j.*.*., 14)
Level 4	(<i>a.b.c.d., 10</i>)	(h.i.m.n., 15)	(w.x.y.z., 11)

Figure 3: Example depicting our one-dimensional algorithm on a stream of IP addresses at byte-wise granularity, where each instance of Space Saving maintains 3 counters. The top grid depicts the state at time T , and the bottom grid depicts the state at time $T+1$, after processing the update $(w.x.y.z., +3)$. The minimum counter for each instance of Space Saving is boldfaced and italicized. If OutputHHH1D is run at time $T+1$ with threshold $\phi N = 12$, the approximate HHHs output would be $h.i.m.n.$, $w.x.y.*.$, $h.i.j.*.$, $a.b.c.*.$, and $q.r.*.*.$

The following lemma is useful in proving that our one-dimensional algorithm satisfies various nice properties.

Lemma 3.1. Define $H_p \subseteq P$ as the set $\{h : h \in P, h \prec p, \nexists h' \in P : h \prec h' \prec p\}$. Then in one dimension, $F_p = f(p) - \sum_{h \in H_p} f(h)$.

Proof. By Definition 2.4,

$$F_p = \sum_{(e \in S) \wedge (e \preceq p) \wedge (e \not\preceq P_p)} f(e) = f(p) - \sum_{(e \in S) \wedge (e \preceq P_p)} f(e).$$

Since the hierarchy is one-dimensional, for each $e \in S$ such that $e \preceq P_p$, there is exactly one $h \in H_p$ such that $e \preceq h$ (otherwise, there would be $h \neq h'$ in H_p such that $h \prec h'$). Thus,

$$\begin{aligned} f(p) - \sum_{(e \in S) \wedge (e \preceq P_p)} f(e) &= f(p) - \sum_{h \in H_p} \sum_{(e \in S) \wedge (e \preceq h)} f(e) \\ &= f(p) - \sum_{h \in H_p} f(h). \end{aligned}$$

□

Theorem 3.2. *Using $O(\frac{H}{\epsilon})$ space, our one-dimensional algorithm satisfies the Accuracy and Coverage requirements of Definition 2.4.*

Proof. By Equation 1, each instance of Space Saving requires $\frac{1}{\epsilon}$ counters, corresponding to $O(\frac{1}{\epsilon})$ space, in order to estimate the unconditioned frequency of each item assigned to it within additive error ϵN . Consequently, the Accuracy requirement is satisfied using $O(\frac{H}{\epsilon})$ space in total.

To prove coverage, we first show by induction that $s_p = \sum_{h \in H_p} f_{\min}(h)$. This is true at level L because in this case $s_p = 0$ and H_p is empty. Suppose the claim is true for all prefixes at level k . Then for p at level $k - 1$,

$$\begin{aligned} s_p &= \sum_{q \in \text{child}(p) \wedge q \in P} f_{\min}(q) + \sum_{q \in \text{child}(p) \wedge q \notin P} s_q \\ &= \sum_{q \in \text{child}(p) \wedge q \in P} f_{\min}(q) + \sum_{q \in \text{child}(p) \wedge q \notin P} \sum_{h \in H_q} f_{\min}(h) \\ &= \sum_{h \in H_p} f_{\min}(h), \end{aligned}$$

where the first equality holds by inspection of Lines 5-9 of the output procedure, and the second equality holds by the inductive hypothesis. This completes the induction.

By Lemma 3.1, $F_p = f(p) - \sum_{h \in H_p} f(h)$

$$\leq f_{\max}(p) - \sum_{h \in H_p} f_{\min}(h) = f_{\max}(p) - s_p,$$

where the inequality holds by the Accuracy guarantees. Coverage follows, since our algorithm is conservative. That is, if item p is not output, then from Line 6 of the output procedure we have $f_{\max}(p) - s_p \leq \phi N$, and we've shown $F_p \leq f_{\max}(p) - s_p$. □

We remark that under realistic assumptions on the data distribution, our algorithm satisfies the Accuracy and Coverage requirements using space $o(\frac{H}{\epsilon})$. Specifically, [2, Theorem 8] shows that, if the tail of the frequency distribution (i.e. the quantity $N^{\text{res}(k)}$ for a certain value of k) is bounded by that of the Zipfian distribution with parameter α , then Space Saving requires space $O(\epsilon^{-\frac{1}{\alpha}})$ to estimate all frequencies within error ϵN . Notice that if the frequency distribution of the stream itself satisfies this “bounded-tail” condition, then the frequency distributions at higher levels of the hierarchy do as well. Hence our algorithm requires only space $O(H\epsilon^{-\frac{1}{\alpha}})$ if the tail of the stream is bounded by that of a Zipfian distribution with parameter α .

Theorem 3.3. *Our one-dimensional algorithm performs each update operation in time $O(H \log \frac{1}{\epsilon})$ in the case of arbitrary updates, and $O(H)$ time in the case of unitary updates. Each output operation takes time $O(\frac{H}{\epsilon})$.*

Proof. The time bound on insertions is trivial, as an insertion operation requires updating H instances of Space Saving. Each update of Space Saving using a min-heap based implementation for arbitrary updates requires time $O(\log m)$, where $m = \frac{1}{\epsilon}$ is the number of counters maintained by each instance of Space Saving. For unitary updates, each insertion to Space Saving can be processed in $O(1)$ time using the *Stream Summary* data structure [18].

To obtain the time bound on output operations, notice that although the pseudocode for procedure OutputHHH1D indicates that we iterate through every possible prefix e , we actually need only iterate over those e tracked by the instance of Space Saving corresponding to e 's label. We may restrict our search to these e because, for any prefix e not tracked by the corresponding Space Saving instance, $f_{\max}(e) \leq \epsilon N < \phi N$, so e cannot be an approximate HHH. There are at most $\frac{H}{\epsilon}$ such e 's because each of the H instances of Space Saving maintains only $\frac{1}{\epsilon}$ counters, and for each e , the GetEstimateSS call in line 5 and all operations in lines 6-9 require $O(1)$ time. The time bound follows. \square

For all prefixes p in the lattice, define the estimated conditioned count of p to be $F'_p := f_{\max}(p) - s_p$. By performing a refined analysis of error propagation, we can bound the number of HHHs output by our one-dimensional algorithm, and use this result to provide Accuracy guarantees on the estimated conditioned counts.

Theorem 3.4. *Let $\epsilon < \frac{\phi}{2}$. The total number of approximate HHHs output by our one-dimensional algorithm is at most $\frac{1}{\phi - 2\epsilon}$. Moreover, the maximum error in the approximate conditioned counts, $F'_p - F_p$, is at most $\frac{1}{\phi - 2\epsilon}\epsilon N$.*

Proof. We first sketch why not too many approximate HHHs are output. A prefix p is output if and only if $F'_p > \phi N$, and $F'_p \geq F_p$. The key observation is that for each approximate HHH $h \in P$ output by our algorithm, h “contributes” error at most ϵN to the estimated conditioned count F'_p of at most one ancestor $p \in P$ of h . Therefore, the total error in the approximate conditioned counts of the output set P is small. Consequently, the sum of the *true* conditioned counts F_p of all $p \in P$ is very close to $\phi N|P|$, implying that $|P|$ cannot be much larger than $\frac{N}{\phi}$ since the stream has length N .

We make this argument precise. We showed in proving Theorem 3.2 that for all p , $s_p = \sum_{h \in H_p} f_{\min}(h)$, so

$$F'_p = f_{\max}(p) - s_p = f_{\max}(p) - \sum_{h \in H_p} f_{\min}(h). \quad (2)$$

Combining Lemma 3.1 and Equation 2, we see that

$$\begin{aligned} F'_p - F_p &= (f_{\max}(p) - \sum_{h \in H_p} f_{\min}(h)) - (f(p) - \sum_{h \in H_p} f(h)) = \\ &= (f_{\max}(p) - f(p)) + (\sum_{h \in H_p} f(h) - f_{\min}(h)). \end{aligned}$$

To show that the sum of the *true* conditioned counts F_p of all $p \in P$ is very close to $\phi N|P|$, we use

$$\sum_{p \in P} F_p = \sum_{p \in P} F'_p - \sum_{p \in P} (F'_p - F_p)$$

$$\geq |P|\phi N - \sum_{p \in P} (f_{\max}(p) - f(p)) - \sum_{p \in P} \left(\sum_{h \in H_p} f(h) - f_{\min}(h) \right).$$

By the Accuracy guarantees, $\sum_{p \in P} (f_{\max}(p) - f(p))$ is at most $|P|\epsilon N$. To bound $\sum_{p \in P} (\sum_{h \in H_p} f(h) - f_{\min}(h))$, we observe that for any item $h \in P$, $h \in H_p$ for at most one ancestor $p \in P$ (because in one dimension, if $h \prec p$ and $h \prec p'$ for distinct $p, p' \in P$, then either $p \prec p'$ or $p' \prec p$, contradicting the fact that $h \in H_p$ and $h \in H_{p'}$). Combining this fact with the Accuracy guarantees, we again obtain an upper bound of $|P|\epsilon N$. In summary, we have shown that

$$\sum_{p \in P} F_p \geq |P|\phi N - 2\epsilon|P|N = |P|(\phi - 2\epsilon)N.$$

Since the total length of the stream is N , and in one dimension each fully specified item contributes its count to F_p for at most one p , it follows that $\sum_{p \in P} F_p \leq N$ and hence $|P| \leq \frac{1}{\phi - 2\epsilon}$ as claimed.

Lastly, we bound the maximum error $F'_p - F_p$ in any estimated conditioned count. We showed that

$$F'_p - F_p = (f_{\max}(p) - f(p)) + \left(\sum_{h \in H_p} f(h) - f_{\min}(h) \right),$$

which, by the Accuracy guarantees, is at most $\epsilon N + |H_p|\epsilon N \leq \epsilon N + (|P| - 1)\epsilon N \leq \frac{\epsilon}{\phi - 2\epsilon}N$, as claimed. \square

The upper bound on output size provided in Theorem 3.4 is very nearly tight, as there may be $\frac{1}{\phi}$ exact heavy hitters. For example, with realistic values of $\phi = .01$ and $\epsilon = .001$, Theorem 3 yields an upper bound of 102.

4 Two-Dimensional Hierarchies

In moving from one to multiple dimensions, only the output procedure must change. In one dimension, discounting items that were already output as HHHs was simple. There was no double-counting involved, since no two children of an item p had common descendants. To deal with the double-counting, we use the principle of inclusion-exclusion in a manner similar to [8] and [7].

At a high level, our two-dimensional output procedure works as follows. As before, we start at the bottom of the lattice, and compute HHHs one level at a time. For any node p , we have to estimate the conditioned count for p by discounting the counts of items that are already output as HHHs. However, Lemma 3.1 no longer holds: it is not necessarily true that $F_p = f_{\max}(p) - \sum_{q \in H_p} f(q)$ in two or more dimensions, because for fully specified items that have two or more ancestors H_p , we have subtracted their count multiple times. Our algorithm compensates by adding these counts back into the sum.

Before formally presenting our two-dimensional algorithm, we need the following theorem. Let $\text{glb}(h, h')$ denote the greatest lower bound of h and h' , that is, the unique common descendant q of h and h' satisfying $\forall p : (q \preceq p) \wedge (p \preceq h) \wedge (p \preceq h') \implies p = q$. In the case where h and h' have no common descendants, then we treat $\text{glb}(h, h')$ as the “trivial item” which has count 0.

Theorem 4.1. *In two dimensions, let T_p be the set of all q expressible as the greatest lower bound of two distinct elements of H_p , but not of 3 or more distinct elements in H_p . Then*

$$F_p = f(p) - \sum_{q \in H_p} f(q) + \sum_{q \in T_p} f(q).$$

The proof appears in Appendix A.

Below, we give pseudocode for our two-dimensional output procedure. We compute estimated conditioned counts $F'_p = f_{\max}(p) - \sum_{h_1 \in H_p} f_{\min}(h_1) + \sum_{q \in T_p} f_{\max}(q)$. As in the one-dimensional case, the Accuracy guarantees of the algorithm follow immediately from those of Space Saving. Coverage requirements are satisfied by combining Theorem 4.1 with the Accuracy guarantees.

Our two-dimensional algorithm performs each insert operation in $O(H \log \frac{1}{\epsilon})$ time under arbitrary updates, and $O(H)$ time under unitary updates, just as in the one-dimensional case. Although the output operation is considerably more expensive in the multi-dimensional case, experimental results indicate that this operation is not prohibitive in practice (see Section 6).

OUTPUTHHH2D(threshold ϕ)

```

1   $P = \emptyset$ 
2  for level  $l=L$  downto 0
3      for each item  $p$  at level  $l$ 
4          Let  $n$  be the lattice node that  $p$  belongs to
5           $(f_{\min}(p), f_{\max}(p)) = \text{GetEstimateSS}(SS(n), p)$ 
6           $F'_p = f_{\max}(p)$ 
7           $H_p = \{h \in P \text{ such that } \nexists h' \in P : h \prec h' \prec p\}$ 
8          for each  $h \in H_p$ 
9               $F'_p = F'_p - f_{\min}(h)$ 
10         for each pair of distinct elements  $h, h'$  in  $H_p$ 
11              $q = \text{glb}(h, h')$ 
12             if  $\nexists h_3 \neq h, h'$  in  $H_p$  s.t.  $q \preceq h_3$ 
13                  $F'_p = F'_p + f_{\max}(q)$ 
14         if  $F'_p \geq \phi N$ 
15              $P = P \cup \{p\}$ 
16         print( $p, f_{\min}(p), f_{\max}(p)$ )
```

Using Theorem 4.1, we obtain a non-trivial upper bound on the number of HHHs output by our two-dimensional algorithm. The proof is in Appendix A.

Theorem 4.2. *Let $A = 1 + \min(h_1, h_2)$, where h_i is the depth of dimension i of the lattice. For small enough ϵ , the number of approximate HHHs output by our two-dimensional algorithm is at most*

$$\frac{2}{A\epsilon} \left(\phi - (1 + A)\epsilon - \sqrt{(\phi - (1 + A)\epsilon)^2 - A^2\epsilon} \right).$$

The error guarantee obtained from Theorem 4.2 appears messy, but yields useful bounds in many realistic settings. For example, for IP addresses at byte-wise granularity, $A = 5$. Plugging in $\phi = .1$, $\epsilon = 10^{-4}$ yields $|P| \leq 53$, which is very close to the maximum number of exact HHHs: $A/\phi = 50$. As further examples, setting $\phi = .05$ and $\epsilon = 10^{-5}$ yields a bound of $|P| \leq 102$, and setting $\phi = .01$ and $\epsilon = 10^{-6}$ yields a bound of $|P| \leq 536$, both of which are reasonably close to $\frac{A}{\phi}$. Of course, the bound of Theorem 4.2 should not be viewed as tight in practice, but rather as a worst-case guarantee on output size.

Higher Dimensions. In higher dimensions, we can again keep one instance of Space Saving at each node of the hierarchy to compute estimates $f_{\min}(p)$ and $f_{\max}(p)$ of the unconditioned count of each prefix p . We need only modify the Output procedure to conservatively estimate the *conditioned* count of each prefix.

We can show that the natural generalization of Theorem 4.1 does not hold in three dimensions. However, we can compute estimated conditioned sublattice counts F'_p as

$$F'_p = f(p) - \sum_{h \in H_p} f_{\min}(h) + \sum_{(h \in H_p, h' \in H_p) \wedge q = \text{glb}(h, h')} f_{\max}(q).$$

Inclusion-exclusion implies that, in any dimension, $F_p \leq F'_p$, and hence by outputting p if $F'_p \geq \phi N$ we can satisfy Coverage.

5 Extensions

Our algorithms are easily adopted to distributed or parallel settings, and can be efficiently implemented in commodity hardware such as ternary content addressable memories.

Distributed Implementation. In many practical scenarios a data stream is distributed across several locations rather than localized at a central node (see, e.g., [16, 21]). For example, multiple sensors may be distributed across a network. We extend our algorithms to this setting.

Multiple independent instances of Space Saving can be merged to obtain a single summary of the concatenation of the distributed data streams with only a constant factor loss in accuracy, as shown in [2]. We use this form of their result:

Theorem 5.1. ([2, Theorem 11], simplified statement): *Given summaries of k distributed data streams produced by k instances of Space Saving each with $\frac{1}{\epsilon}$ counters, a summary of the concatenated stream can be obtained such that the error in any estimated frequency is at most $3\epsilon N$, where N is the length of the concatenated stream.*

To handle k distributed data streams, we may simply run one instance of our algorithm independently on each stream (with $\frac{3}{\epsilon}$ counters each), and afterward, for each node in the lattice, merge all k corresponding instances of Space Saving into a single instance. After the merge, we have a single instance of Space Saving for each node in the lattice that has essentially the same error guarantees (up to a small constant factor) as a centralized implementation. Our output procedure is exactly as in the centralized implementation.

Parallel Implementation. In all of our algorithms, the update operation involves updating a number of independent Space Saving instances. It is therefore trivial to parallelize this algorithm. We have parallelized this algorithm using OpenMP. Our limited experiments show essentially linear speedup, up to the point where we reach the limitation of the shared memory constraint.

TCAM Implementation. Recently, there has been an effort to develop network algorithms that utilize Ternary Content Addressable Memories, or *TCAMs*, to process streaming queries faster. TCAMs are specialized, widely deployed hardware that support constant-time queries for a bit vector within a database of ternary vectors, where every bit position represents 0, 1 or *. The * is a wild card bit matching either a 0 or a 1. In any query, if there is one or more match, the address of the highest-priority match is returned. Previous work describes a TCAM-based implementation of Space Saving for unitary updates, and shows experimentally that it is several times faster than software solutions [1].

Since our algorithms require the maintenance of H independent instances of Space Saving, it is easy to see that our algorithms can be implemented given access to H separate TCAMs, each requiring just a few KBs of memory. With more effort, we can devise implementations of our algorithms that use just a single commodity TCAM. Commodity TCAMs can store hundreds of

thousands or millions of data entries [1], and therefore a single TCAM can store tens of instances of Space Saving even when $\epsilon = .0001$.

Our simplest TCAM-based implementation takes advantage of the fact that TCAMs have *extra bits*. A typical TCAM has a width of 144 symbols allotted for each entry in the database, and this typically leaves several dozen unused symbols for each entry. The implementation of Space Saving of [1] uses extra bits to store frequencies, but we can use additional unused bits to identify the instance of Space Saving associated with each item in the database.

For illustration, consider the one-dimensional byte-wise IP hierarchy. We associate two extra bits with each entry in the database: 00 will correspond to the top-most level of the hierarchy, 01 to the second level, 10 to the third, and 11 to the fourth. Then we treat each IP address $a.b.c.d$ as four separate searches: $a.b.c.d.00$, $a.b.c.*.01$, $a.b.*.*.10$, and $a.*.*.*.11$, thereby updating each ancestor of $a.b.c.d$ in turn. The TCAM needs to store the smallest counter for each of the four Space Saving instances, and otherwise the TCAM-based implementation from [1] is easily modified to handle multiple instances of Space Saving on a single TCAM.

Alternatively, we could compute approximate unconditioned counts by keeping a *single* instance of Space Saving with (item, mask) pairs as keys, rather than H separate instances of Space Saving. It is clear that this approach still satisfies the Accuracy guarantees for each prefix, and has the advantage of only having to store the smallest counter for a single instance of Space Saving.

Sliding Windows and Streams with Deletions. Our algorithms as described only work for insert-only streams, due to our choice of Space Saving as our heavy hitter algorithm. However, the accuracy and coverage guarantees of our HHH algorithms still hold even if we replace Space Saving with other heavy hitter algorithms. This is because our proofs of accuracy and coverage applied the inclusion-exclusion principle to express conditioned counts in terms of unconditioned counts, and then used the fact that our heavy hitter algorithm provides accurate estimates on the unconditioned counts; this analysis is independent of the heavy hitter algorithm used. Hence we can extend our results to additional scenarios by using other algorithms.

For example, it may be desirable to compute HHHs over only a sliding window of the last n items seen in the stream. [14] presents a deterministic algorithm for computing ϵ -approximate heavy hitters over sliding windows using $O(1/\epsilon)$ space. Thus, by replacing Space Saving with this algorithm, we obtain an algorithm that computes approximate HHHs over sliding windows using space $O(H/\epsilon)$, which asymptotically matches the space usage of our algorithm. However, it appears this algorithm is markedly slower and less space-efficient in practice.

Similarly, many *sketch-based* heavy hitter algorithms such as that of [9] can compute ϵ -approximate heavy hitters, even in the presence of deletions, using space $O(\frac{1}{\epsilon} \log N)$. By replacing Space Saving with such a sketch-based algorithm, we obtain a HHH algorithm using space $O(\frac{H}{\epsilon} \log N)$ that can handle streams with deletions. (As noted previously, this variation was mentioned in [8].)

6 Experimental Results

We have implemented two versions of our algorithm in **C** and tested it using GCC version 4.1.2 on a host with four single-core 64-bit AMD Opteron 850 processors each running at 2.4GHz with a 1MB cache and 8GB of shared memory. The first version – termed **hhh** below – uses a heap-based implementation of Space Saving that can handle arbitrary updates, while the second version – termed **unitary** below – uses the *Stream Summary* data structure and can only handle unitary updates. Both versions use an off-the-shelf implementation from [4] for Space Saving; further optimizations, as well as different tradeoffs between time and space, may be possible by modifying the off-the-shelf implementation. We have used a real packet trace from www.caida.org [3] in all experiments below. (We have tried other traces to confirm that these results are demonstrative.

Note that all of our graphs are in color and may not display well in grayscale.) Throughout our experiments, all algorithms define the frequency of an IP address or an IP address pair to be the number of packets associated with that item, as opposed to the number of bytes of raw data. This ensures that all algorithms (including **unitary**) process exactly the same updates. Consequently, the stream length N in all of our experiments refers to the number of packets in the stream (i.e. the prefix of the packet trace [3] that we used).

We tested our algorithms at both byte-wise and bit-wise granularities in one and two dimensions. Bit-wise hierarchies are more expensive to handle, as H , the number of nodes in the lattice structure implied by the hierarchy, becomes much larger. However, it may be useful to track approximate HHHs at bit-wise granularity in many realistic situations. For example, a single entity might control a subnet of IP addresses spanning just a few bits rather than an entire byte. However, we observed similar (relative) behavior between all algorithms at both bit-wise and byte-wise granularity, and thus we display results only for byte-wise hierarchies for succinctness.

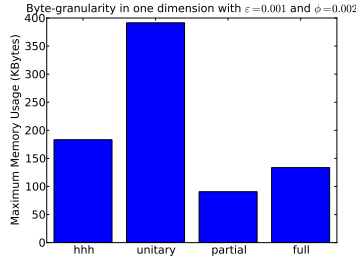
For comparison we also implemented the full and partial ancestry algorithms from [8], labeled **full** and **partial** respectively. We compare the algorithms’ performance in several respects: time and memory usage, the size of the output set, and the accuracy of the unconditioned count estimates. Our algorithm performs at least as well as the other two in terms of output size and accuracy. Except for extremely small values of ϵ (less than about .0001), which correspond to extremely high accuracy guarantees, our two-dimensional algorithm is also significantly faster (more than three times faster for some parameter settings of high practical interest). Our one-dimensional algorithm is also faster than its competitors for values of ϵ greater than about .01, and competitive across all values of ϵ . Our algorithm uses slightly more memory than its competitors. Below, we discuss each aspect separately.

In summary, our one-dimensional algorithm is competitive in practice with existing solutions, and possesses other desirable properties that existing solutions lack, such as improved simplicity and ease of implementation, improved worst-case guarantees, and the ability to preallocate space even without knowledge of the stream length. Our two-dimensional algorithm possesses all of the same desirable properties, and is also significantly faster than existing solutions for parameter values of primary practical interest. The primary disadvantage of our algorithms is slightly increased space usage.

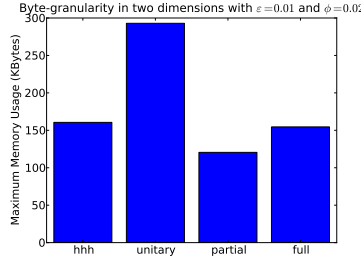
All of our implementations are available online at [19].

Memory. Both versions of our algorithm use more memory than **full** and **partial**. The difference between **hhh**, **partial**, and **full** is a small constant factor; **unitary** uses about twice as much space as **hhh**. The largest difference in space usage appears in one dimension, as shown in Figure 4a. The difference is much smaller in two dimensions, as shown in Figure 4b. In both cases, the better space usage of **partial** comes at the cost of significantly decreased accuracy and increased output size, as discussed below. We conclude that in situations where the decreased accuracy of **partial** cannot be tolerated, the memory usage of our algorithms is not a major disadvantage, as **hhh** and **full** have similar memory requirements, especially in two dimensions.

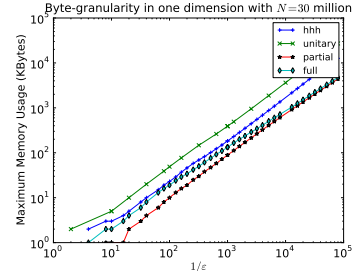
Ideally, we would be able to present our results with the independent variable a programmer-level object such as memory usage, rather than the error-parameter ϵ . In practice, a programmer may be allowed at most 1 MB of space to deploy an HHH algorithm on a network sensor, and have to optimize speed and accuracy subject to this constraint. But while the mapping between ϵ and memory usage is straightforward for our algorithm (the Space Saving implementation uses 36 bytes per counter, and we use a fixed $\frac{H}{\epsilon}$ counters), this mapping is less clear for **partial** and **full**, as their space usage is data dependent, with counters added and pruned over the course of execution. Figures 4c and 4d show the empirical mapping between space usage and ϵ for a fixed stream length



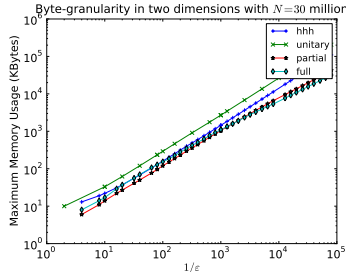
(a) Maximum memory usage in one dimension over all stream lengths N . For **hhh** and **unitary**, space usage does *not* depend on N ; space usage only varied with N for **partial** and **full**.



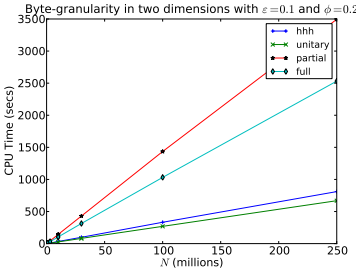
(b) Maximum memory usage in two dimensions over all stream lengths N . For **hhh** and **unitary**, space usage does *not* depend on N ; space usage only varied with N for **partial** and **full**.



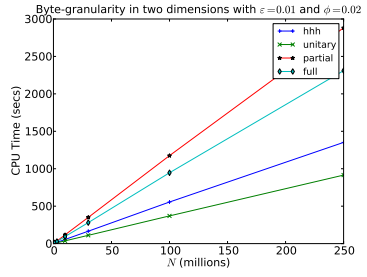
(c) Memory usage in one dimension for fixed stream length.



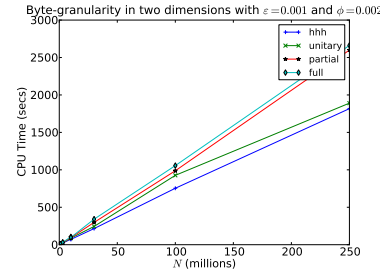
(d) Memory usage in two dimensions for fixed stream length.



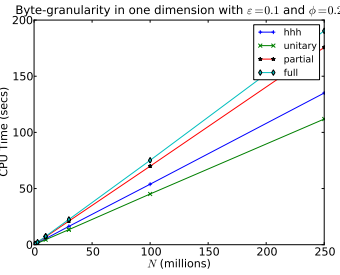
(e) Speed comparison in two dimensions with high ϵ .



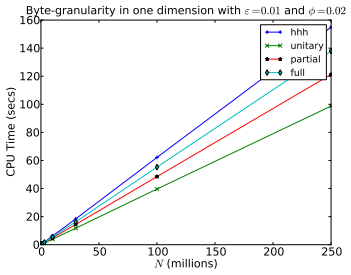
(f) Speed comparison in two dimensions with medium ϵ .



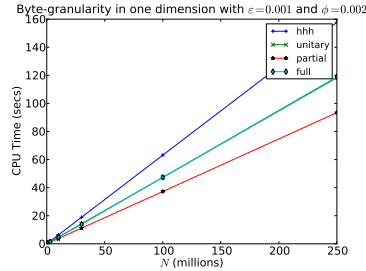
(g) Speed comparison in two dimensions with low ϵ .



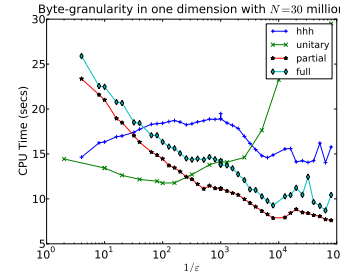
(h) Speed comparison in one dimension with high ϵ .



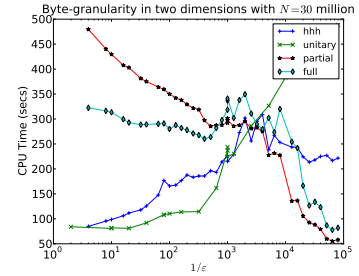
(i) Speed comparison in one dimension with medium ϵ .



(j) Speed comparison in one dimension with low ϵ .

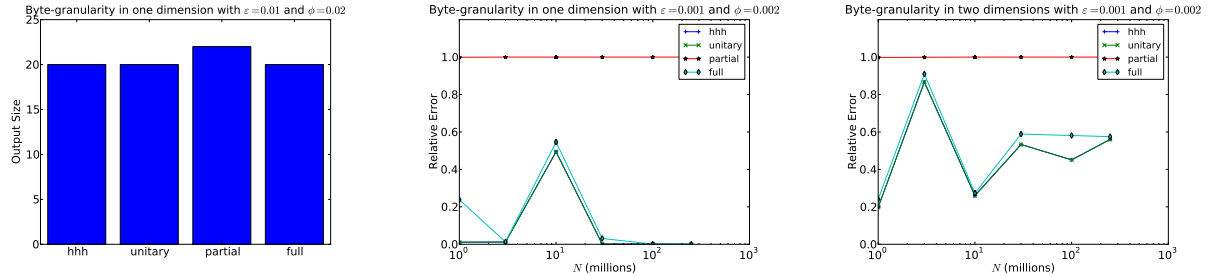


(k) Speed comparison in one dimension for fixed stream length.



(l) Speed comparison in two dimensions for fixed stream length.

Figure 4: Memory and speed comparisons.



(a) Output size comparison. For each algorithm, we display the maximum output size over all streams tested (for the given setting of ϵ and ϕ). (b) Accuracy comparison in one dimension. (c) Accuracy comparison in two dimensions.

Figure 5: Accuracy and output size comparisons.

of $N = 30$ million with one- and two-dimensional bitwise hierarchies. This setting highlights the importance of our improved worst-case space bounds, even though our algorithm uses slightly more space in practice. It can be imperative to *guarantee* assigned memory will not be exceeded, and our algorithm allows a more aggressive choice of error parameter while maintaining a worst-case guarantee.

We emphasize that we did *not* attempt to optimize memory usage for our algorithms using characteristics of the data, as suggested in Theorem 3.2. It is therefore likely that our algorithms can function with less memory than **partial** and **full** in many practical settings.

Note that the running time and memory usage are independent of ϕ , as ϕ only affects the output stage, which we have not included in our measurements as the resources consumed by this stage were negligible.

Time. We observe that in both one and two dimensions, both **unitary** and **hhh** are faster than **partial** and **full** except for extremely small values of ϵ . The speed of each algorithm for each setting of ϵ is illustrated for a fixed stream length of $N = 30$ million in Figures 4k and 4l; our algorithms are fastest in one dimension for ϵ greater than about .01 and in two dimensions for ϵ greater than about .0001.

We show how runtime grows with stream length for fixed values of ϵ in Figures 4e-4j. For concreteness, on a stream with $N = 250$ million, **unitary** processes about 2.2 million updates per second in one dimension at a byte-wise granularity when $\epsilon = .1$, while **hhh** processes 1.85 million, **partial** 1.3 million, and **full** processes 1.4 million. Here, N corresponds to the number of packets (not weighted by size), and the updates per second statistic specifies the number of packets processed per second by our implementation. In two dimensions for $\epsilon = .1$, **unitary** processes over 370,000 updates per second and **hhh** processes 300,000, while **partial** processes 71,000, and **full** processes about 100,000. Thus, our algorithms ran more than three times faster than **partial** and **full** for this particular setting of parameters.

Output Size. The output size and accuracy are measures of the quality of output and depend on the value of ϕ . All three algorithms produce a near-optimal output size, with **partial** consistently outputting the largest sets. The largest difference observed is shown in Figure 5a.

Accuracy. We define the relative error of the output to be $\max_{p \in \text{output}} \frac{f_{\max}(p) - f_{\min}(p)}{\epsilon N}$. Clearly, the relative error is between 0 and 1, because of the accuracy guarantees of our algorithms. We find that the relative error can vary significantly for all algorithms, but our algorithm uniformly performs best. The relative error of the partial ancestry algorithm is often close to the theoretical

upper bound of 1, making it by far the least accurate of the algorithms tested.

TCAM Simulations. We simulated our TCAM-conscious implementation of our algorithm on the same packet traces as above, in order to estimate the number of TCAM operations our implementation requires per packet processed. [1] experimentally demonstrates that TCAM READ, WRITE, and SEARCH operations all take roughly the same amount of time. Thus, we counted the total number of READ, WRITE, and SEARCH operations our TCAM implementation required, without distinguishing between the three. We found that for one-dimensional IP addresses at byte-wise granularity, each packet required about 14 TCAM operations on average, or 2.8 TCAM operations per instance of Space Saving maintained by our algorithm. This is slightly better than the worst-case behavior of the implementation of [1], which requires up to 4 TCAM operations per update. Our two-dimensional algorithm at byte-wise granularity requires about 65 TCAM operations per packet; since the two-dimensional algorithm maintains 25 instances of Space Saving, this translates to only 2.6 TCAM operations per instance of Space Saving. We attribute this improvement in TCAM operations per instance of Space Saving to the fact that the frequency distribution at high nodes in the two-dimensional lattice is highly non-uniform.

7 Conclusion

The trend in the literature on the approximate HHH problem has been towards increasingly complicated algorithms. In this work, we present what is perhaps the simplest algorithm for HHHs in arbitrary dimension, and demonstrate that it is superior to the existing standard in many respects, and competitive in all others. We believe our algorithm offers the best tradeoff between simplicity and performance.

Acknowledgements. We thank Elaine Angelino, Richard Bates, and Graham Cormode for helpful discussions. Michael Mitzenmacher is supported by NSF grants CNS-0721491, CCF-0915922, and IIS-0964473, and grants from Cisco, Inc., Google, and Yahoo!. Justin Thaler is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

- [1] N. Bandi, A. Metwally, D. Agrawal, and A. El Abbadi. Fast data stream algorithms using associative memories. In *Proc. of the 2007 ACM SIGMOD Intl. Conf. on Management of Data*, pages 247–256.
- [2] R. Berinde, P. Indyk, G. Cormode, and M. J. Strauss. Space-optimal heavy hitters with strong error bounds. *ACM Trans. Database Syst.*, 35:26:1–26:28, October 2010.
- [3] W. Colby, E. Aben, K. Claffy, and D. Andersen. The CAIDA anonymized Internet traces. At <https://data.caida.org/datasets/oc48/oc48-original/20030115/1hour/20030115-100000-1-anon.pcap.gz>.
- [4] G. Cormode and M. Hadjieleftheriou. Finding frequent items in data streams: Source code. <http://www.research.att.com/~marioh/frequent-items.html>.
- [5] G. Cormode and M. Hadjieleftheriou. Finding frequent items in data streams. *Proc. VLDB Endow.*, 1:1530–1541, August 2008.

- [6] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava. Finding hierarchical heavy hitters in data streams. In *Proc. of the 29th Intl. Conf. on Very Large Data Bases*, pages 464–475, 2003.
- [7] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava. Diamond in the rough: Finding hierarchical heavy hitters in multi-dimensional data. In *Proc. of the 2004 ACM SIGMOD Intl. Conf. on Management of Data*, pages 155–166, 2004.
- [8] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava. Finding hierarchical heavy hitters in streaming data. *ACM Trans. Knowl. Discov. Data*, 1:2:1–2:48, February 2008.
- [9] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.
- [10] C. Estan and G. Magin. Interactive traffic analysis and visualization with Wisconsin Netpy. In *Proc. of the 19th Conf. on Large Installation System Administration*, pages 177–184, 2005.
- [11] C. Estan, S. Savage, and G. Varghese. Automatically inferring patterns of resource consumption in network traffic. In *Proc. of the 2003 ACM SIGCOMM Conf.*, pages 137–148, 2003.
- [12] J. Hershberger, N. Shrivastava, S. Suri, and C. D. Tóth. Space complexity of hierarchical heavy hitters in multi-dimensional data streams. In *Proc. of the Twenty-Fourth ACM Symp. on Principles of Database Systems*, pages 338–347, 2005.
- [13] L. Jose, M. Yu, and J. Rexford. Online measurement of large traffic aggregates on commodity switches. In *Proc. of the 11th USENIX Conf. on Hot Topics in Management of Internet, Cloud, Enterprise Networks and Services*, 2011.
- [14] L.K. Lee and H.F. Ting. A simpler and more efficient deterministic scheme for finding frequent items over sliding windows. In *Proc. of the Twenty-Fifth ACM Symp. on Principles of Database Systems*, pages 290–297, 2006.
- [15] Y. Lin and H. Liu. Separator: Sifting hierarchical heavy hitters accurately from data streams. In *Proc. of the 3rd Intl. Conf. on Advanced Data Mining and Applications*, pages 170–182, 2007.
- [16] A. Manjhi, V. Shkapenyuk, K. Dhamdhere, and C. Olston. Finding (recently) frequent items in distributed data streams. In *Proc. of the 21st Intl. Conf. on Data Engineering*, pages 767–778, 2005.
- [17] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. of the 28th Intl. Conf. on Very Large Data Bases*, pages 346–357, 2002.
- [18] A. Metwally, D. Agrawal, and A. Abbadi. Efficient computation of frequent and top- k elements in data streams. In T. Eiter and L. Libkin, editors, *Database Theory - ICDT 2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 398–412.
- [19] M. Mitzenmacher, T. Steinke, and J. Thaler. Hierarchical heavy hitters with the space saving algorithm: Source code. <http://people.seas.harvard.edu/~tsteinke/hhh/>.
- [20] S. Muthukrishnan. Data streams: algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 2005.

- [21] C. Olston, J. Jiang, and J. Widom. Adaptive filters for continuous queries over distributed data streams. In *Proc. of the 2003 ACM SIGMOD Intl. Conf. on Management of Data*, pages 563–574, 2003. ACM.
- [22] V. Sekar, N. Duffield, O. Spatscheck, J. van der Merwe, and H. Zhang. LADS: large-scale automated DDoS detection system. In *Proc. of the 2005 USENIX Annual Technical Conf.*, pages 171–184, 2006.
- [23] P. Truong and F. Guillemin. Identification of heavyweight address prefix pairs in IP traffic. In *Proc. of the 21st Intl. Teletraffic Congress*, 2009.
- [24] P. Turan. On an extremal problem in graph theory. *Matematiko Fizicki Lapok*, 48:436–352, 1941.
- [25] Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund. Online identification of hierarchical heavy hitters: algorithms, evaluation, and applications. In *Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement*, pages 101–114, 2004.

A Proof of Theorems

Proof of Theorem 4.1. First note that it is possible, using the inclusion-exclusion principle, to show that

$$\begin{aligned}
 F_p = & f(p) - \sum_{h \in H_p} f(h) + \sum_{(h, h' \in H_p) \wedge q = \text{glb}(h, h')} f(q) \\
 & - \sum_{(h, h', h'' \in H_p) \wedge q = \text{glb}(h, h', h''')} f(q) + \dots
 \end{aligned}$$

We claim that for all u expressible as the greatest lower bound of more than two elements of H_p , the total contribution of $f(u)$ to the above sum is 0. Indeed, suppose that $u = (u_1, u_2)$ is a descendant from exactly m such elements, h_1, h_2, \dots, h_m in H_p . Since $u \prec h_\alpha$ for α in $\{1, \dots, m\}$ these m heavy hitter elements can be written as $h_1 = (P_{i_1}u_1, P_{j_1}u_2)$, $h_2 = (P_{i_2}u_1, P_{j_2}u_2)$, \dots , $h_m = (P_{i_m}u_1, P_{j_m}u_2)$, where $(P_i u_1, P_j u_2)$ denotes the element obtained from (u_1, u_2) by generalizing i times on the first dimension and j times on the second dimension. Renumbering if necessary, assume the nodes are sorted on the generality of their first component so that $i_1 < i_2 < \dots < i_m$. It is clear that there are no equalities in the sequence because if $i_\alpha = i_\beta$ then either $h_\alpha \preceq h_\beta$ or $h_\beta \preceq h_\alpha$ which contradicts that these are from H_p . When the corresponding relationships between the second components are examined it can be seen that the increasing sort on the first component forces a decreasing order on the second component $j_1 > j_2 > \dots > j_m$ (since if $\alpha < \beta$ and $j_\alpha \leq j_\beta$ then because $i_\alpha < i_\beta$ the contradiction $h_\alpha \prec h_\beta$ is reached). Thus the m elements are in a linear structure with endpoints h_1 and h_m . Clearly the first component of u is the first component of h_1 and similarly the second component of u is the second component of h_m . With this in hand it is clear that u is the greatest lower bound of any subgroup of the m elements that includes h_1 and h_m . There are $\binom{m-2}{k}$ ways to pick k middle terms and thus there are $\binom{m-2}{k}$ ways in which node u appears as the greatest lower bound of $k+2$ elements from H_p . Returning to the sum

$$\begin{aligned}
 F_p = & f(p) - \sum_{h \in H_p} f(h) + \sum_{(h, h' \in H_p) \wedge q = \text{glb}(h, h')} f(q) \\
 & - \sum_{(h, h', h'' \in H_p) \wedge q = \text{glb}(h, h', h''')} f(q) + \dots
 \end{aligned}$$

it is now clear that $f(u)$ will appear once in the sum over pairs, $m-2$ times in the sum over triples, and, in general, $\binom{m-2}{k}$ times in the sum over groups of size $k+2$. When combined with the sign structure in the sum this gives a resulting contribution from u of

$$f(u) \sum_{j=0}^{m-2} (-1)^j \binom{m-2}{j} = f(u)(1-1)^{m-2} = 0.$$

Thus in the two dimensional case

$$F_p = f(p) - \sum_{h_1 \in H_p} f(h_1) + \sum_{q \in T_p} f(q)$$

as claimed. \square

Proof of Theorem 4.2. The proof will closely parallel that of Theorem 3.4. We bound the total error in the estimated conditioned counts, aggregated over all $p \in P$, and this will imply that the sum of the true conditioned counts of all $p \in P$ is large. Hence there cannot be too many approximate HHHs output.

We showed in Theorem 4.1 that

$$F_p = f(p) - \sum_{h_1 \in H_p} f(h_1) + \sum_{q \in T_p} f(q).$$

Therefore,

$$\begin{aligned} F'_p - F_p &= (f_{\max}(p) - f(p)) + \sum_{h_1 \in H_p} (f(h_1) - f_{\min}(h_1)) \\ &\quad + \sum_{q \in T_p} (f_{\max}(q) - f(q)). \end{aligned}$$

Our goal is to show that the sum of the true conditioned counts of all $p \in P$ is large by bounding the total error in the estimated conditioned counts, aggregated over all $p \in P$. To this end, consider the sum

$$\begin{aligned} \sum_{p \in P} F_p &= \sum_{p \in P} F'_p - \sum_{p \in P} (F'_p - F_p) \geq |P|\phi N - \sum_{p \in P} (F'_p - F_p) \\ &= |P|\phi N - \sum_{p \in P} (f_{\max}(p) - f(p)) - \\ &\quad \sum_{p \in P} \sum_{h_1 \in H_p} (f(h_1) - f_{\min}(h_1)) - \sum_{p \in P} \sum_{q \in T_p} (f_{\max}(q) - f(q)). \end{aligned}$$

We refer to the second term on the right hand side of the last expression, $\sum_{p \in P} (f_{\max}(p) - f(p))$, as “Term-Two error”, the third term, $\sum_{p \in P} \sum_{h_1 \in H_p} (f(h_1) - f_{\min}(h_1))$, as “Term-Three” error, and the fourth term, $\sum_{p \in P} \sum_{q \in T_p} (f_{\max}(q) - f(q))$ as “Term-Four error”. By the Accuracy guarantees, it is immediate that the Term-Two error is bounded above by $|P|\epsilon N$.

In order to bound Term-Three error, we must briefly introduce the notion of comparable items in a lattice. Two elements x and y are *comparable* under the \preceq relation if the label of y is less than or equal to that of x on every attribute. Let A be the size of the largest antichain in the lattice, that is, the maximum size of any subset of prefixes such that any two items in the subset are incomparable. It was shown in [8] that $A = 1 + \min(h_1, h_2)$. We show that $\sum_{p \in P} |H_p| \leq A|P|$; it then follows by the Accuracy guarantees that the Term-Three error is bounded above by $|P|A\epsilon N$.

To this end, for any $h \in P$, consider the set $B_h = \{p \in P : h \in H_p\}$. We claim that $|B_h| \leq A$, since all the items in B_h must be incomparable. Indeed, suppose $p, q \in B_h$ and the label of q is less than the label of p on both attributes. Then $h \prec q \prec p$, so by definition of H_p , $h \notin H_p$, which is a contradiction. Thus, $\sum_{p \in P} |H_p| = \sum_{h \in P} |B_h| \leq A|P|$.

Finally, we may bound the Term-Four error by $\frac{AP^2}{4}\epsilon N$. This will clearly follow from the Accuracy guarantees if we can bound $\sum_{p \in P} |T_p|$ by $\frac{A|P|^2}{4}$. To this end, for each $p \in P$ let G_p be a graph on $|H_p|$ vertices, where edge $(h_1, h_2) \in E(G_p)$ if and only if $\text{glb}(h_1, h_2) \in T_p$. It is clear that $|T_p| = |E(G_p)|$. We claim G is a triangle-free graph – it then follows by Turan's theorem [24] that $|T_p| \leq \frac{|H_p|^2}{4}$. For three distinct vertices $h_1, h_2, h_3 \in H_p$, let $u_1 = \text{glb}(h_1, h_2)$, $u_2 = \text{glb}(h_2, h_3)$ and $u_3 = \text{glb}(h_1, h_3)$. We show that if u_1, u_2 and u_3 are all in T_p , then for at least one i , u_i is a descendant of h_1, h_2 , and h_3 , contradicting $u_i \in T_p$.

Write $h_i = (h_{i,1}, h_{i,2})$ for each $i \in \{1, 2, 3\}$. By assumption (h_i, h_j) share a common descendant for any pair (i, j) , so we may assume (renumbering if necessary) that $h_{1,1} \prec h_{2,1} \prec h_{3,1}$ as one-dimensional objects. The remainder of the proof now closely parallels that of Theorem 4.1. It is clear that there are no equalities in the sequence because if $h_{i,1} = h_{j,1}$ then either $h_{i,1} \preceq h_{j,1}$ or $h_{j,1} \preceq h_{i,1}$ which contradicts that these are from H_p . For the same reason, it can be seen that the increasing sort on the first component forces a decreasing order on the second component, i.e., $h_{3,2} \prec h_{2,2} \prec h_{1,2}$. Consequently, $u_3 = \text{glb}(h_1, h_3) = (h_{1,1}, h_{3,2})$ is a descendant of h_1, h_2 , and h_3 , contradicting $u_3 \in T_p$.

So we have shown that $|T_p| \leq \frac{|H_p|^2}{4}$. Since $\sum_{p \in P} |H_p| \leq A|P|$, and trivially $|H_p| \leq |P|$ for all $p \in P$, Holder's Inequality implies that $\sum_{p \in P} |T_p| \leq \sum_{p \in P} \frac{|H_p|^2}{4} \leq A|P|^2/4$.

Thus, we have shown that

$$\sum_{p \in P} F_p \geq |P|\phi N - |P|\epsilon N - A|P|\epsilon N - \frac{A|P|^2}{4}\epsilon N.$$

Now note that $AN \geq \sum_{p \in P} F_p$, because each fully-specified item e can only contribute to the true conditioned counts of incomparable approximate HHHs. For if e contributes to the conditioned count of both p and q then $e \preceq p \wedge e \not\preceq P_p$ and $e \preceq q \wedge e \not\preceq P_q$. If p and q are comparable, then this implies either $q \preceq p$ or $p \preceq q$, contradicting the fact that $e \not\preceq P_p$ and $e \not\preceq P_q$. Thus, we see that $AN \geq (\phi N - (A+1)\epsilon N)|P| - \frac{A|P|^2}{4}\epsilon N$.

Dividing through by N and subtracting A from both sides yields

$$0 \geq -A + (\phi - (A+1)\epsilon)|P| - \frac{A\epsilon}{4}|P|^2.$$

Using the quadratic equation, this holds if and only if

$$|P| \leq -2 \frac{(1+A)\epsilon - \phi + \sqrt{(\phi - (1+A)\epsilon)^2 - A^2\epsilon}}{A\epsilon}$$

or

$$|P| \geq -2 \frac{(1+A)\epsilon - \phi - \sqrt{(\phi - (1+A)\epsilon)^2 - A^2\epsilon}}{A\epsilon},$$

and we can rule out the latter case for small enough ϵ via trivial upper bounds on $|P|$ such as $|P| \leq \frac{H}{\phi}$. \square